Report of the October 2002 Meeting
of the
Science Archive Working Group

The SAWG held their second meeting on October 31-November 1, 2002 at NASA HQ, with the following members present: Julian Borrill, Joel Bregman (Chair), Roger Brissenden, Damian Christian, Menas Kafatos, Bill Oegerle (Deputy Chair), Tom McGlynn, Sally Oey, Rick White, along with the NASA HQ personnel Paul Hertz, Jefferey Hayes, and Joe Bredekamp.

NED, ADS, the ADEC, and Virtual Observatory Activities

There was a update by Bob Hanisch who is project manager of the NSF ITR project on Virtual Observatory (VO) infrastructure, which has just concluded its first of five years of support. Basic goals in regards to various standards are being achieved and there appears to be international cooperation between various efforts around the world, which generally focus on different infrastructure components. At the January 2003 AAS meeting, the NSF ITR program will have three demonstrations of VO applications as applied to real data.

The SAWG heard reports from the NED and ADS, both on improvements of the sites as well as user statistics. One surprising result is that there are approximately an order of magnitude more unique users than there are working astronomers (e.g., 60,000 unique users per month for ADS) and since the identity of the majority of the users are unknown, it was suggested that NED and ADS come to a better understanding of their user base. We should be aware if our archival sites are being used fruitfully for educational instruction, and if so, whether we can further assist in these activities.

These and other NASA archival data centers have continued to take steps for their interoperability, which is one of the key elements of the VO concept. They are enthusiastic about advancing the VO capabilities with their NSF counterparts but are limited from doing so by insufficient resources. Consequently, the SAWG suggests that it is an appropriate time for the archival centers to increase their interoperability in order to meet strategic goals and to prepare for NASA participation in the anticipated VO. In particular, this development of VO-related activities should be considered along the lines of a NASA Project that will support the primary goals of the SEUS and OS roadmaps, in concert with the data that would be collected from the envisioned missions. Project Requirements should flow from these considerations, and there should be a well-defined set of data standards, goals, milestones, staffing levels, and budgets along a three-year timetable with a nominal start date in FY04. A "white paper" would be the result of this planning. This is envisioned as a modest NASA-only program of limited scope in which the staffing and budget models should be described for both an optimum and a minimal program.

The SAWG suggests that the natural group to plan these activities be a committee comprised of the ADEC, with input from scientists involved in the NSF component of the VO, and possibly members of the user community who actively use the archive sites. We suggest that the ADEC appoint one of its members to take the lead in producing this document. This planning activity builds upon previous efforts that produced the science definition document developed last year. This committee should move forward in a timely fashion to produce the white paper by March 2003, in time for review in the next SAWG meeting, which we anticipate will be in April 2003.

Data Analysis Tools

The diversity and lack of compatibility of existing software systems is a serious hindrance to effective research in astronomy. Requiring users to learn distinct analysis environments to analyze data from different instruments and missions is tedious and unproductive. We urge support for the development of common analysis frameworks. The intent of the frameworks need not be to restrict users to either a single implementation of algorithms or to promote a specific analysis environment. Experience has shown that especially for the lower level analysis tools, mission specific software is essential. However a common framework might define standards and protocols for the interfaces between analysis environments and tools such that software developed in one environment could be easily used in another. Such frameworks will be even more critical in the Virtual Observatory environment. It is the intent of the ADEC to address this issue of consolidating certain software environments and we strongly endorse this effort, which we intend to follow in future meetings.

A related issue is that NASA currently funds the development of software through its Applied Systems and Information Research Program (ASIRP) program, and it once funded some software development through the ADP program. Some of these programs were for the development of tools that were to become of general use, and while a few became valuable, many did not. A concern is that the development of tools frequently needs to occur in relationship to an existing package (e.g., FTOOLS) in order to be useful and this should be taken into consideration for the funding of future programs.

MO&DA Issues

There was discussion about the MO&DA grant programs, dealing with whether funding levels were appropriate. As a general guideline, the committee felt that once the over-subscription rate for funds rises above 3:1, excellent science programs are rejected and NASA is being shortchanged on the scientific return from its missions. One of the worst cases is the ATP, where the current over-subscription rate is 7.5:1, but there are plans to correct this in the coming year by the infusion of new funds, which should reduce this ratio to 3:1 - 4:1. The ADP and LTSA programs have an over-subscription rate approaching 5:1 and the committee felt that additional funds would be of benefit here as well. For the LTSA program, we applaud the philosophy of "the best science per dollar" as the primary criteria for proposal selection. Also, the distinction of "Junior" and "Senior" LTSA proposals might be eliminated, as it does not seem to add significantly to the program selection process.

The general belief that a 3:1 over-subscription rate is a sensible target is based upon the experience of various committee members, but it does not represent a scientific study. A more justifiable over-subscription target should be obtained by more quantitative means. For example, a review committee might occasionally be convened two years after the award date has ended to grade the effectiveness and value of the approved programs. Then, one could judge whether the programs with the lowest scores during the initial review process were significantly less valuable than the more highly rated programs. This would "close the loop" on the proposal process and should be of value to the MO&DA staff in evaluating their programs.

GLAST Proprietary Data Rights

The GLAST instruments have a broad field of view, making it difficult to allocate observations from one point source to a PI. Instead, the GLAST team has developed a proprietary rights plan in which there will be a proprietary period for the scientific "idea" rather than for the photons, which are to become available immediately after adequate processing. While the SAWG commends the GLAST team for "thinking outside the box", we have serious concerns about the feasibility and desirability of the proposed proprietary rights policy for the GLAST data. It is unclear how the notion of proprietary "ideas" can be defined or enforced, and it seems antithetical to the scientific enterprise. While the proposed three month proprietary period is nominally short, the clock starts only at the end of a potentially multi-year investigation, so that the effective proprietary periods can be many years. Defining the details of tracking the usability of each photon for given investigations – and providing that information to the community – is daunting.

The SAWG recognizes the difficulty in defining proprietary data for large field-of-view missions such as GLAST. The GLAST developers may wish to consider whether it may be possible to dispense with a proprietary period altogether for this mission. Influencing the observing schedule, receiving funding, and having directed SSC support might be adequate rewards for guest observers without requiring this complex notion of proprietary ideas. In general, rights based upon access to resources, rather than to ideas, seem more consistent with the spirit of scientific inquiry and certainly appear more practical.

GALEX

David Schiminovich gave a nice overview of the GALEX mission, whose principal goal is to study the star formation history of the universe through ultraviolet radiation from hot stars. GALEX is currently slated for a February 2003 launch. Schiminovich provided us with a timely update on the data processing and archiving plans that appear very thoughtful, especially for a cost-constrained SMEX budget. The committee is pleased to see the re-use of existing archive software developed for the Sloan Digital Sky Survey, and that meta-data are being prepared in tables that are compatible with emerging NVO standards.

The committee would like to make several suggestions for improvements to the archived data products and support for the Guest Observer (GO) program (referred to as the Associate Investigator Program, AIP, by DS):

(1) The current plans are to archive only the reconstructed 2D sky maps, 1D extracted spectra and catalogs in MAST. The SAWG highly recommends that the raw, photon list data also be archived at MAST along with sufficient pointing control data to allow a user to reconstruct sky maps and spectra. Although raw data will eventually be deep-archived at the NSSDC, it is especially useful to researchers to be able to access the data from MAST during the active mission phase. In order to be of maximum use, the raw photon list data should also be stored in a format that is usable by standard FITS software tools (i.e. in FITS binary table format, for example).

(2) The first release of data to the public is not scheduled until June 2004, which means that the GOs will not have seen real flight data from GALEX before the proposal deadlines for cycle 1 or possibly even cycle 2 (assuming an extended mission). This is a very undesirable situation that we think can be easily solved by providing a very early release of a small, representative

sample of data. This would provide users with a look at the characteristics of the data, and not overburden the GTO team with having to verify the large volume of data currently planned for the first data release (5% of the sky survey). We think it is especially desirable to provide this quick release data in time for GOs to see it before the cycle 1 GO proposal deadline. Finally, we note that early access to even a small set of data will allow GOs to develop data analysis tools that are not currently being made available by the GALEX project.

(3) We understand that a first draft of the Project Data Management Plan (PDMP) has just been issued. Although not explicitly addressed at the SAWG meeting, we would like to understand the "proprietary rights" policy for GALEX data for both the GTO and GO teams, and we hope that those issues are addressed in the PDMP.

The SAWG looks forward to successful orbital operations of GALEX next year.

The Formation of User Committees

The SAWG was asked to consider the best time to create a User Committee for a new project. During discussion it was agreed that User Committees are important to ensure an independent and adequate science voice during the life of a mission. In addition to providing a user view, they are well suited to review the first AO at about a year before launch, and this provides one possible milestone for the start of the Committee. However, there is also a need to provide user input during the development phase to ensure that data tools, early policy development and archive plans (for example) are suitable for users. Code S may wish to consider forming a smaller core Committee for providing such input at appropriate project milestones during development, and then expanding the Committee to full size for the first AO. This is especially important for missions where the external users have major scientific interests that are not well represented by scientists on the development team.

Funding of FFRDCs to Conduct Ground-Based Observations

A problem exists in that researchers at FFRDCs cannot apply to the NSF Astronomy division for funding astronomy-related research projects. In some cases, these research programs involve ground-based astronomy projects that are naturally funded by the NSF. This restriction only applies to the astronomy division of the NSF, as researchers at FFRDCs can apply for and hold grants from other parts of the NSF, such as the physics division. We think this problem would be best solved not by having NASA fund these ground-based astronomy projects, but rather by having the NSF fund ground-based research at FFRDCs. The SAWG strongly encourages a dialogue between NASA and the NSF to discuss the funding of ground-based astronomy in a consistent manner that is beneficial to all of astronomy.

Initiation of the LAMBDA Archive

The SAWG received a status report on the Legacy Archive for Microwave Background Data Analysis (LAMBDA) from its director Gary Hinshaw. We note that LAMBDA is not expected to be a new long-lasting archive (like MAST), since the next large mission, PLANCK, is to be archived at the US PLANCK data center at IPAC. Understandably LAMBDA's current activities are very focused on its preparations for ingesting, serving and supporting the MAP data, and this should be their primary short-term goal. Aside from the activities of serving the

MAP data, the LAMBDA team should formulate clear plans for a 1 year and 3 year time frame that involve the development of the archive's overall data holdings and their associated analysis software and servicing facilities. The LAMBDA team might consider forming a users committee to assist in planning. LAMBDA should also look to the other GSFC archival groups located in HEASARC and NSSDC for advice on planning the archive and assisting users.

Several suggestions were made as to the activities that LAMBDA might take on, beyond its MAP archival efforts. For example, LAMBDA could offer to be a repository for data and software from other CMB groups, which might be extremely valuable. There was some concern that considerable effort is needed to properly add sub-orbital data to this archive (with meta-data and documentation) and the various missions are not funded to carry out this additional work. Consequently, the cost and benefit of adding such data sets would need to be addressed before making this a LAMBDA priority.

The NASA Information Technology Vision

The new Chief Information Officer for NASA is Paul Strassman, who spent much of his career at Xerox and is very experienced in large computer systems and issues of data storage. He discussed a variety of computer issues at NASA, relating to cost, uniformity of installed software, long-term archiving of data, and security, where there have been some publicized security breaches, and there are many unpublicized attacks per day (such as at Ames Research Center, the location for one of the thirteen root servers for the WWW). The SAWG appreciates the need for security and a certain amount of software uniformity within NASA, so we discussed whether these goals would conflict with our focus: making certain that outside users have convenient access to archival data. Mr. Strassman was very concerned and assured us that he understands the issue of archival access and that the work of NASA scientists will not be compromised by his policies. He stressed that he is readily available and any issues should be sent directly to him. Another concern is that some policies will have costs associated with them, in the form of new software, hardware, or manpower. If such new costs place severe demands on the limited resources of an archive site, we hope that the CIO will provide some assistance.

The next meeting of the SAWG is expected to take place in April 2003 and we welcome suggestions from the SEUS and the OS for future topics to be addressed.